Physical Modeling of a Bag Knot in a Robot Learning System

Uri Kartoun, Amir Shapiro, Helman Stern, and Yael Edan, *Members*, *IEEE*

***Abstract*—This paper presents a physical model developed to find the directions of forces and moments required to open a plastic bag - which forces will contribute toward opening the knot and which forces will lock it further. The analysis is part of the implementation of a *Q(λ)*-learning algorithm on a robot system. The learning task is to let a fixed-arm robot observe the position of a plastic bag located on a platform, grasp it, and learn how to shake out its contents in minimum time. The physical model proves that the learned optimal bag shaking policy is consistent with the physical model and shows that there were no subjective influences. Experimental results show that the learned policy actually converged to the best policy.**

***Index Terms*—**Robot learning, reinforcement learning, intelligent robots, robot kinematics

I. Introduction

To expand the use of robots in everyday tasks they must be able to perform in unpredictable and continuously changing environments. Since it is impossible to model all environments and task conditions in a rigorous enough manner, robots must learn independently how to respond to the world and how the world responds to actions they take.

One approach to robot learning is reinforcement learning (RL) (*e.g.*, [1]-[9]). In RL the robot receives positive/negative rewards from the environment indicating how well it performs the required task. The robot learning goal is to optimize system responses by maximizing a reward function. The advantages of RL are that a detailed model of the problem and a training set are not required and actions are modeled with non-deterministic outcomes.

The RL-based learning algorithm Q-learning [10], and its variation $Q(\lambda)$ [11], an incremental multi-step Q-learning algorithm that combines one-step Q-learning with eligibility traces, have been used in many robotic applications [*e.g.*, [1], [4], [5], [12]). In an effort to test a learning algorithm on a real robot platform, the $Q(\lambda)$-learning algorithm was tested with a fixed-arm six degrees of freedom Motoman UP6 robot manipulating a bag containing objects. The task was to empty the bag of its contents which is initially placed on a flat table. For a robot to empty the bag, one side of its gripper must slide under the bottom of the bag, grasp it and lift it to a "shake starting position" vertically over the table. The bag is held upright by the robot from its bottom so that its opening, initially closed by a knot, is facing down. The learning task is to observe the position of the bag and learn how to shake the knot loose so that all of its contents fall out in minimum time. Several assumptions are made: (i) a plastic bag has already been recognized; (ii) the bag has been grasped to its shaking starting position over the table, and (iii) the knot on the bag is such that it can be opened by shaking. The reward function is updated when a falling object event is detected by a digital scale placed under the table. To interpret the results and to show that there were no subjective influences, an analytical model of the bag and its knot was developed to validate the learned policy. The model has the ability to explain the result of the robot system described in the paper, using Newton's fundamental laws of physics.

Similar tasks described in the literature refer mainly to tying and untying tasks that involve strings or cords. See for example, knot-tying tasks addressed in [13] and [14]. In [15] a topological model is developed, to recognize knot structures in order to overcome the robots' difficulty in manipulating deformable objects such as ropes. The paper describes a recognition method to distinguish the structure of a rope using two knot invariants. A new motion planner for manipulating deformable linear objects, such as ropes, cables and sutures is presented in [16]. The planner was implemented in simulation using two cooperating robot arms to tie various knots around simple static objects. Most of this work is concerned with the robot manipulator grasping the rope itself for the purpose of tying or untying a knot. Our model is different in that the robot manipulator does not grasp a string or rope, but instead grasps a plastic bag, which is closed by tying the plastic handles of the bag into a knot tying. The robot manipulator tries to open the bag by shaking loose the knot in lieu of grasping the knot itself.

After reviewing the Q and $Q(\lambda)$ algorithms in Section II, the $Q(\lambda)$ implementation is detailed in Section III. Section IV presents the results of applying the algorithm on the system. A physical modeling of the system is presented in Section V. Discussion and conclusions are provided in Sections VI and VII, respectively.

II. Q And $Q(\lambda)$-Learning Review

The basic assumption in Markov Decision Processes is that the probability distribution of any state s_{t+1} occupied by an

Manuscript received April 2, 2008; revised September 13, 2008. This work was partially supported by the Paul Ivanier Center for Robotics Research and Production Management, the Rabbi W. Gunther Plaut Chair in Manufacturing Engineering and the Pearlstone Center for Aeronautical Engineering, Ben-Gurion University of the Negev.

U. Kartoun, H. Stern, and Y. Edan are with the Department of Industrial Engineering and Management; A. Shapiro is with the Department of Mechanical Engineering. All authors are from the Ben-Gurion University of the Negev, Beer-Sheva 84105, ISRAEL (phone: +972-8-6461434; fax: +972-8-6472958; e-mail: kartoun@bgu.ac.il; ashapiro@bgu.ac.il; helman@bgu.ac.il; yael@bgu.ac.il).

agent is a function of its last state and action: $s_{t+1} = f(s_t, a_t)$ where $s_t \in S$ and $a_t \in A$ are the state and action indices at time step t, respectively. The sets S and A represent the action and state spaces [17].

In Q-learning, a specific Markov system, the system estimates the optimal action-value function directly and then uses it to derive a control policy using the local greedy strategy [10]. It is stated in [5] "Q-learning can learn a policy without any prior knowledge of the reward structure or a transition model." Q-learning is thus referred to as a "model-free approach" where Q values can be calculated directly from the elementary rewards observed. Q is the system's estimate of the optimal action-value function [18]. At each time step t, the agent visits state s_t and selects an action a_t. Then it receives from the process the reinforcement $r(s_t, a_t) \in R$ and observes the next state s_{t+1}. The procedure continues by updating the action value $Q(s_t, a_t)$ according to (1) which describes a Q-learning one step.

$$Q_{t+1}(s_t, a_t) = (1-\alpha)Q_t(s_t, a_t) + \alpha[r(s_t, a_t) + \gamma \hat{V}_t(s_{t+1})] \quad (1)$$

In (1), $\hat{V}_t(s_{t+1}) = \max_{a_t \in A}[Q_t(s_{t+1}, a_t)]$ is the current estimate of the optimal expected cost $V^*(s_{t+1})$ and α is the learning rate which controls how much weight is given to the immediate reward, as opposed to the old Q estimate. The process repeats until a stopping criterion is met. The greedy action $\arg\max_{a_t \in A}[Q_t(s_{t+1}, a_t)]$ is the best the agent performs when at state s_{t+1}. For the initial stages of the learning process, however, actions are chosen randomly to encourage exploration of the environment. Convergence to the optimal value function is guaranteed under some reasonable conditions (bounded rewards and a learning rate is in the range of zero to one) [19] by iteratively applying (1), [18]. To boost the slow learning of the Q algorithm [21], a multi-step tracing mechanism, the eligibility trace, is used [1]. This improved learning method is called $Q(\lambda)$ and is faster than standard Q [20].

III. Bag Shaking Experiment with a Fixed-Arm Robot

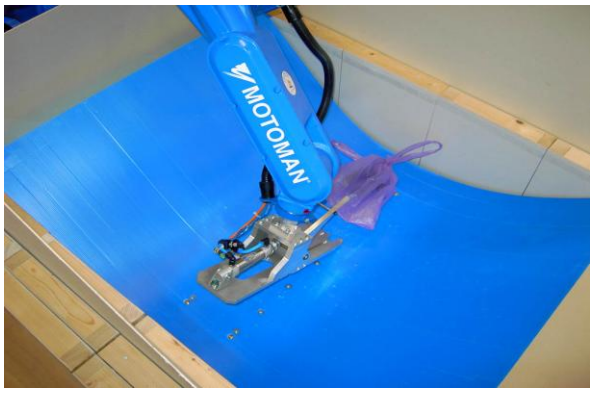
A. Task Definition

The experiment utilizes a Motoman UP-6 fixed-arm robot positioned over a table surface. The learning task is to observe the position of a plastic bag located on a table surface, grasp it with a fixed-arm robot, and learn how to shake out its contents in minimum time using $Q(\lambda)$-learning (Figs. 1 and 2). It is assumed that a plastic bag contains a number of identical objects with known weights.[1] In the experiments performed there were five identical screws. The bag is placed on a table with its opening initially closed by the knot. For a robot to empty the contents of a bag, one side of its gripper must slide under the bottom of the bag (the part opposite the knot), grasp it and lift it to a "shake starting position" vertically over the table. It is assumed that the type of bag and the location of the opening are known. The type of bag grasp operation was determined by a bag classification algorithm achieved by finding optimal features representing a bag using support vector machines (SVMs) [22]. A digital scale placed under the table surface was used to automatically measure rewards.

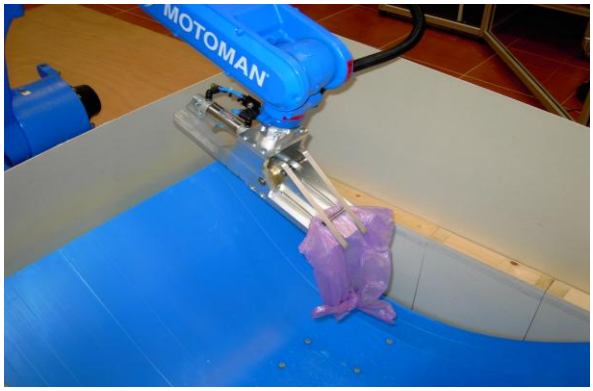
The robot has no a-priori knowledge regarding the most efficient shaking policy for any given plastic bag, but it learns this information from interaction with the environment. In the experiments performed, the first (out of 50) learning episode consists of a random shaking policy over the robot's X, Y, and Z axes. For the experiment, robot states pertain to its gripper location in a three-dimensional grid (Fig. 3). The performance of the task is a function of a set of actions, $a_t \in A$, for each physical state, $s_t \in S$, of the system.

An action, a_t, consists of a robot movement from a state point, s_t, along a single coordinate direction (X, Y, Z). The Y axis which is parallel to the bag's handles is defined as the horizontal shaking axis, *i.e.*, actions are performed in parallel to the horizon (left to right). At the X axis (forward to backward), actions are performed in parallel to a horizon perpendicular to the Y axis. Over the Z axis, actions are performed vertically (up to down). It is assumed that the bag has been grasped and moved to a central point centered 50 cm above a table surface. Further, it is assumed that the bag's knot can be opened by shaking. The robot starts a shaking policy from this $s_{(Center)_t}$ state. From $s_{(Center)_t}$ it can move in the direction of any of the three coordinates reaching any of the 18 possible states (six possible states for each axis). The adjacent state distances in any axis are 30 mm. From any robot state other than $s_{(Center)_t}$, the robot is limited to either symmetrical actions or returning to the center position.

[1] The assumption of equal weights for the items in the bag was made in order to let the robot be rewarded based on the precise number of objects dropped in a particular time. To let the learning task handle items of unequal weight, it would have been necessary to measure the total weight of the bag and all items. Then, during a learning episode, when the digital scale measures such a weight, the learning episode is accomplished.



(a) Plastic bag grasp operation



(b) Robot grasped the plastic bag and moved it to a central point above a table surface. The bag is held upright by the robot gripper and has a hole in the bottom which is initially closed by the knot.

Fig. 1. Experimental setup

(a) Plastic bag and five screws

(b) Closed plastic bag with screws

Fig. 2. Plastic bag with objects

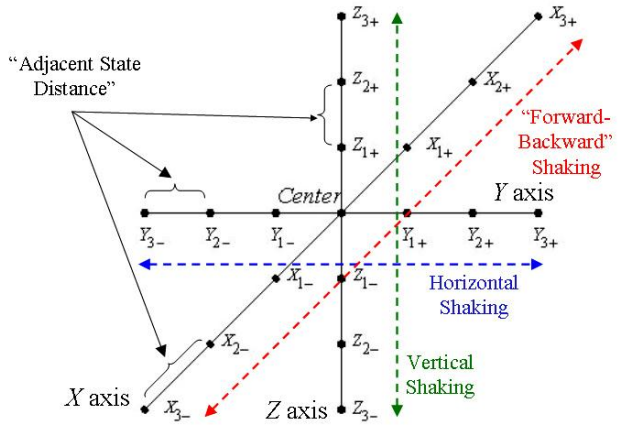


Fig. 3. Motoman UP-6 fixed-arm robot state-space

To evaluate the system performance two measures were calculated after each learning episode: (i) the time it took the robot to empty the contents of the bag, or the time limit whichever is smaller (see T below), and (ii) the reward.

B. *Experiments*

The performance of $Q(\lambda)$ was tested for 50 learning episodes. An event-based weight history was obtained from the digital scale and was used to measure the rewards automatically. The reward function for learning episode n is denoted as R_n as shown in (2).

$$R_n = \sum_{j=0}^{T} \frac{(W_j - W_{j-1})/w}{t_j} \qquad (2)$$

Here W_j is the current weight of all items that were shaken from the bag, measured by a digital scale located under the table surface at time t_j (increments of 0.25 second). If at time t_j one or more objects fall from the bag, dividing the weight differences by t_j effectively increases the reward for items that fall early. Other parameters include w, the weight of one object (a constant value), and W_{j-1}, the weight measured by the scale during the previous time (for j = 1, $W_{j-1} = 0$). *T=min{Fixed Horizon Time, Amount of Time when all Objects Fell}* is the time of shaking (one learning episode), where the *Fixed Horizon Time* is measured as the time it takes the robot to perform a pre-defined number of actions and was set to 100. If no objects fell after 100 actions, it is assumed that the bag is closed and therefore no items fell. This implies that T is calculated as the amount of time it took the robot to perform 100 actions (this time varies from one learning episode to another because different policies may be taken in different learning episodes). The *Amount of Time when all Objects Fell* is the amount of time till the bag is empty. The value $(W_j - W_{j-1})/w$ represents the number of objects that fell at time t_j. Note, if no objects fall at time t_j the numerator equals 0, and the reward is not increased. The robot starts its first learning episode by performing a random shaking policy over the X, Y, and Z axes. The default speeds and adjacent state distance were set to 1000 mm/s and 30 mm respectively for all axes.

IV. Experimental Results

The average time to empty the contents of the bag using $Q(\lambda)$-learning over 50 learning episodes was 10.4 seconds (excluding grasping and lifting). Performance times and rewards are as shown in Figs. 4 and 5, respectively (in the first four learning episodes no items fell from the bag resulting a zero reward).

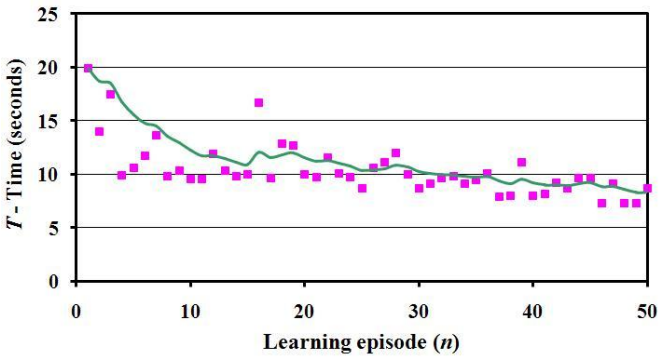


Fig. 4. Performance times for events-based reward

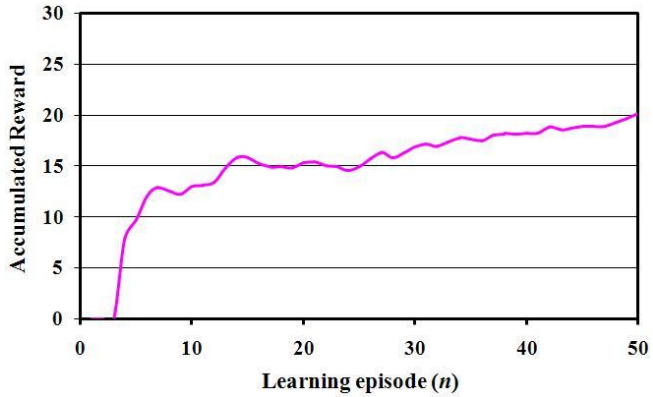


Fig. 5. Reward performance

Intuitively, vertical shaking should work best, but results indicated otherwise, with shaking most of the time over the horizontal Y axis and little activity over the X axis (and no Z axis actions). One possible reason for favoring the Y axis may be the type of knot holding the plastic bag closed; pulling it sideways loosens the knot faster. This is corroborated by a kinematic knot model discussed in the next section.

V. Physical Modeling of a Plastic Bag Knot

A. Overview

To interpret the results and to show that there were no subjective influences, a physical model of the opening of a plastic bag knot by a robot was developed. The model explains the results described in this paper as well as previous experiments performed [7], [23]. It showed that it was worthwhile to open the bag using a continuous shaking/motion from locations as far as possible from the center of the horizontal Y axis.[2] Ideally, the robot arm should be accelerated to match or closely match the gravitational acceleration downwards and should be oscillated over the Y axis to overcome most of the friction forces.

B. 2D Analysis

Analysis includes the best movements for opening the knot while the bag contains only one rigid body (Fig. 6).

[2] Acceleration is constant, but over time position is quadratic, and therefore we need as much distance as possible.

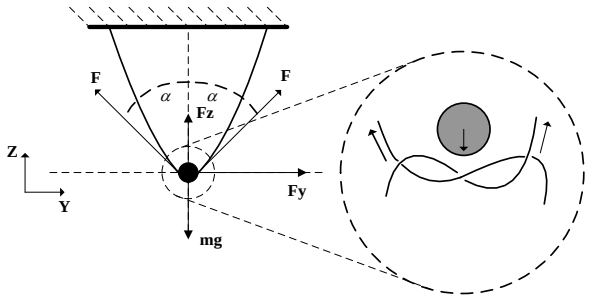


Fig. 6. Two dimensional one object forces (for static and dynamic cases)

Static Case

For the static equilibrium case (Fig. 6), it is assumed that the mass of the bag is negligible with respect to the weight of the object and to the friction forces in the knot. The sum of the forces in the Z direction is:

$$2F\cos\alpha = mg \quad (3)$$

$$F = mg/(2\cos\alpha) \quad (4)$$

where F is the plastic bag tension and α is the angle between the vertical axis and the force vector F. At the contact between the two bag handles F serves as the tangential friction force. Therefore, to open the bag knot by sliding, the two handles on each other (5) shall hold. This yields to (6),

$$F = mg/2\cos\alpha > \mu mg \quad (5)$$

$$\mu < 1/(2\cos\alpha) \quad (6)$$

for the limit on the friction coefficient. This means that when the bag is hanged down, a friction coefficient of at least 0.5 is required to keep it closed and a higher friction coefficient to keep it closed as the ends are pulled apart further.

Dynamic Case

For the dynamic case, when the robot shakes a bag with a knot aligned with the Y axis (Fig. 7), (7) expresses Newton's law of the object along the Y direction:

$$\Sigma F_y = F_y = F\sin\alpha + ma_y \quad (7)$$

where F_y is a force activated on the bag by the object and a_y is the object acceleration. Now we write Newton's law in the Z direction:

$$\sum F_z = \max\left[ma_z - mg, 0\right] \quad (8)$$

where 0 means detachment of the object from the bag. A condition that the knot will be open is (9):

$$\sum F_y > \mu \sum F_z \quad (9)$$

Based on (7) (8) and (9), (10) is organized:

$$F\sin\alpha + ma_y > \mu\max\left[ma_z - mg, 0\right] \quad (10)$$

where both a_y and a_z are controllable by commanding different bag accelerations.

For opening the bag knot, it is desired that the expression $F\sin\alpha + ma_y$ will be as large as possible and the expression $\mu\max[ma_z - mg, 0]$ will be as small as possible. Thereby, it is desired to increase both a_y and a_z. a_z is desired to be increased to $a_z = -g$, then the right hand side in (10) will be equal to zero. The reason for doing that is due to the vertical

force that contributes toward increasing the normal load and thus increasing the friction against opening the bag. On the other hand, accelerating along the Y axis will increase the left hand side of (10) and therefore will act to open the bag. However, a conflict arises when a force is activated over the Y axis for a certain amount of time since the value $\sin\alpha$ is decreased over time. This explains why it is desired to change the side of activating forces over the Y axis of the bag, *i.e.*, to shake it. Another reason for shaking might be that maintaining acceleration in a constant direction results in ever increasing speeds which are not maintainable. In conclusion, it is desirable to shake the bag a little up and down over the Z axis to compensate over the gravitational forces that lock the bag (this decreases the friction between the bag and object) and shake it a lot over the Y axis to slide the bag handles and open the knot.

C. 3D Analysis

Now we move to the case where the knot is no longer aligned with the Y axis, and it is not flat anymore, *i.e.*, each handle is a wide triangular-shaped stripe stretched away from the knot center. Fig. 7 shows the *X*-*Y* plane where the bold dot represents the center of the knot of length 2*l*. The two dotted areas shown in Fig. 7 are three dimensional sectors. Each sector is spanned by two forces, F_u and F_v derived from the robot shaking activity. These forces are tangent to the bag curvature. The vectors F_L and F_R are the left and right tension vectors over the bag knot, each making an angle of θ between F_u and F_v. A three dimensional representation is shown in Fig. 8.

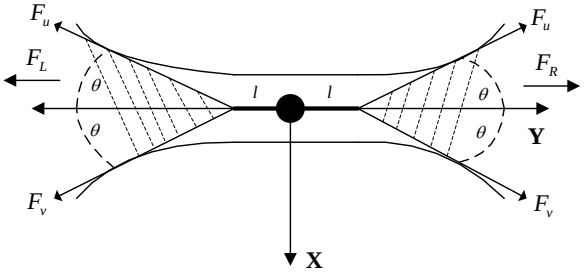


Fig. 7. Top view of a 3D one object and part of the plastic bag bottom

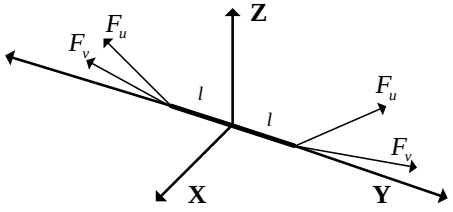


Fig. 8. Three dimensional plastic bag axes

Based on Figs. 7 and 8, equations (11) and (12) are expressed:

$$\mathbb{F}_L = \left\{F_L = \gamma_L F_{U_L} + \delta_L F_{V_L}: \quad \gamma_L, \delta_L \geq 0\right\} \tag{11}$$

$$\mathbb{F}_R = \left\{F_R = \gamma_R F_{U_R} + \delta_R F_{V_R}: \quad \gamma_R, \delta_R \geq 0\right\} \tag{12}$$

where $\mathbb{F}_L$ and $\mathbb{F}_R$ are groups of all forces activated over to the left and to the right of the bag knot, respectively. $F_L \in \mathbb{F}_L$ and $F_R \in \mathbb{F}_R$, respectively, are the left and right possible force values activating in the spanning dotted area (Fig. 7). Based on Newton's second law, (13) and (14) are written:

$$F_R + F_L + mg = ma \tag{13}$$

$$x: F_{L_x} + F_{R_x} = ma_x$$

$$y: F_{L_y} + F_{R_y} = ma_y \tag{14}$$

$$z: F_{L_z} + F_{R_z} - mg = ma_z$$

where a_x, a_y, and a_z are the accelerations at the X, Y, and Z axes, g is the gravitational acceleration and mg is the force that the object activates on the bag.[3] Under the constraint for opening the bag:

$$\max\left[\sqrt{F_{R_x}^2 + F_{R_y}^2}, \sqrt{F_{L_x}^2 + F_{L_y}^2}\right] > \mu(mg + ma_z) \tag{15}$$

where $\sqrt{F_{R_x}^2 + F_{R_y}^2}$ and $\sqrt{F_{L_x}^2 + F_{L_y}^2}$ are tangential forces, *i.e.*, components parallel to the bag surface and $\mu(mg + ma_z)$ is the normal force, *i.e.*, a component perpendicular to the surface of the bag. Rearranging (15), yields:

$$\max\left[F_{R_x}^2 + F_{R_y}^2, F_{L_x}^2 + F_{L_y}^2\right] > \mu^2 m^2 (g + a_z)^2 \tag{16}$$

Using (14) and (16), yields:

$$\max\left[F_{R_x}^2 + F_{R_y}^2, (ma_x - F_{R_x})^2 + (ma_y - F_{R_y})^2\right] > \mu^2 m^2 (g + a_z)^2 \tag{17}$$

It is required to accelerate the bag at a direction that is opposite to the direction of the force in order to maximize the left hand side of (17). Further, the right hand side of (17) is required to be minimal as possible and this can be achieved by accelerating the bag downwards. However, accelerating the bag downwards at a high value might cause the object in the bag to collide with the robot's gripper, thereby the value of this acceleration should be bounded by $g - \varepsilon$ where $\varepsilon \to 0$. Ideally, it is desirable to accelerate the robot arm at $a_z = g - \varepsilon$ downwards and to oscillate it over the Y axis; this to overcome of most friction forces. Since acceleration is developed over time, it is worthwhile to open the bag by activating forces continuously while holding locations as far as possible over the Y axis. Further, due to the unique structure of the plastic bag, the knot length 2*l* is not a negligible length and θ is small (Figs. 7 and 8). If the length of *l* was close to 0 then there was no preference to activate any force exists in the sector that is bounded by F_u and F_v. Since *l* is significantly larger than 0 then at the section (+*l*, -*l*) around the bag knot center, it is preferable to activate forces over the Y axis.

VI. Discussion

At first look, the robotic task described in the paper appears simple; to only allow movement on one of three axes, to only allow symmetrical movements in those axes, and to only allow switching between axes at one point. The robot's grasping location and the bag opening orientation are known as well. It is also assumed that the items in the bag have an identical weight. However, while the location of the bag opening is known in advance, learning to find the directions to shake the

[3] It should be noted that Newton's second law as expressed in (13) and (14) applies on the object inside the bag, however based on Newton's third law, equal forces (but with opposite reaction) apply on the bag itself.

bag optimally within minimal time using reinforcement learning by a robot is not a trivial task. Relaxing the assumption on knowing the location of the knot could end up with a different shaking policy. For example if the knot is on the side after grasping instead of on the bottom, one would suspect the shaking would be more on the vertical direction to loosen the knot and more in the horizontal direction in order to cause the items to "jump" out of the bag. In this case the physical model could be extended to take this into account and accelerate the items to the correct direction.

The $Q(\lambda)$ algorithm performs surprisingly well, given the notoriously slow convergence of reinforcement learning algorithms. Another contribution of the paper is the development of a detailed physical model of the process of opening a bag knot, and comparing it with the learned control policy. Although the model was developed for a specific knot type of a plastic bag, we suspect that the type of knot is crucial; different types of bags with different openings and textures would definitely affect the optimal shaking policy that the robot will find and the material of the bag affects the coefficient of friction. Additional concerns that should be considered in the future include the texture of a bag and whether the knot has a zipper or a lace (for example, as in suitcases and backpacks).

VII. Conclusions

A physical model of a plastic bag knot demonstrates the consistency with the learned robot policy to shake out the contents of a bag. Further, it is shown that the optimal bag shaking policy uses a continuous shaking/motion from locations as far as possible from the center of the horizontal robot axis. By applying *Q(λ)*-learning the robot converged to the same policy that was derived from the model. The innovation of the model is its ability to explain the result of a physical learning process (such as the robot system described in the paper) using Newton's fundamental physics laws. The model serves as a test case for the learning. Since learning works well here, we suspect it will work well in more complex systems.

The model provides the directions of forces and moments required to open a plastic bag - which forces will contribute toward opening the knot and which forces will lock it further. This result suggests that both the model and the learning process are valid since they independently converged to the same optimal solution. According to the physical model it is necessary to shake the bag along the axis of the knot. However, when a robot grasps the bag it has no way to know what the actual axis of the knot is as it can be along any direction. Therefore, learning is required to actually find the best direction. In our experiments we intentionally direct the knot axis along the Y axis. Hence, a human could verify that the learning process actually converged to the analytical solution, but the analytical solution itself would not be enough to solve the task.

References

[1] W. Zhu and S. Levinson, "Vision-based reinforcement learning for robot navigation," *in Proc. of the Int. Joint Conf. on Neural Networks*, vol. 2, pp. 1025-1030, Washington D.C., 2001.

[2] V. N. Papudesi and M. Huber, "Learning from reinforcement and advice using composite reward functions," *in Proc. 16th International FLAIRS Conf.*, pp. 361-365, St. Augustine, FL, 2003.

[3] V. N. Papudesi, Y. Wang, M. Huber, and D. J. Cook, "Integrating user commands and autonomous task performance in a reinforcement learning framework," *AAAI Spring Symposium on Human Interaction with Autonomous Systems in Complex Environments*, Stanford University, CA., 2003.

[4] P. Kui-Hong, J. Jun, and K. Jong-Hwan, "Stabilization of biped robot based on two mode *Q*-learning," *in Proc. of the 2nd Int. Conf. on Autonomous Robots and Agents*, pp. 446-451, New Zealand, 2004.

[5] R. Broadbent and T. Peterson, "Robot learning in partially observable, noisy, continuous worlds," *in Proc. of the 2005 IEEE Int. Conf. on Robotics and Automation*, pp. 4386-4393, Barcelona, Spain, 2005.

[6] B. Bakker, V. Zhumatiy, G. Gruener, and J. Schmidhuber, "Quasi-online reinforcement learning for robots," *in Proc. 2006 IEEE Int. Conf. on Robotics and Automation*, pp. 2997-3002, 2006.

[7] U. Kartoun, H. Stern, and Y. Edan, "Human-robot collaborative learning system for inspection," *IEEE Int. Conf. on Systems, Man, and Cybernetics*, pp. 4249-4255, Taipei, Taiwan, October, 2006.

[8] L. Mihalkova and R. Mooney, "Using active relocation to aid reinforcement," *in Proc. 19th Int. FLAIRS Conf. (FLAIRS-2006)*, pp. 580-585, Melbourne Beach, Florida, 2006.

[9] R. Ganesan, T. K. Das, and K. M. Ramachandran, "A multiresolution analysis-assisted reinforcement learning approach to run-by-run control," *IEEE Trans. on Automation Science and Engineering*, vol. 4, no. 2, pp. 182-193, 2007.

[10] C. J. C. H. Watkins, "Learning from Delayed Rewards," Ph.D. dissertation, Psychology Dept., Cambridge University, 1989.

[11] J. Peng and R. Williams, "Incremental multi-step *Q*-learning," *Machine Learning*, vol. 22, no. 1-3, pp. 283-290, 1996.

[12] Y. Dahmani and A. Benyettou, "Seek of an optimal way by *Q*-learning," *Journal of Computer Science*, vol. 1, no. 1, pp. 28-30, 2005.

[13] J. Takamatsu, T. Morita, K. Ogawara, H. Kimura, and K. Ikeuchi, "Representation for knot-tying tasks," *IEEE Trans. on Robotics*, vol. 22, no. 1, pp. 65-78, 2006.

[14] H. Wakamatsu, A. Eiji, and H. Shinichi, "Knotting/unknotting manipulation of deformable linear objects," *International Journal of Robotics Research*, vol. 25, no. 4, pp. 371-395, 2006.

[15] T. Matsuno and T. Fukuda, "Manipulation of flexible rope using topological model based on sensor information," *Int. Conf. on Intelligent Robots and Systems*, pp. 2638-2643, 2006.

[16] M. Saha and P. Isto, "Motion planning for robotic manipulation of deformable linear objects," *Int. Conf. on Intelligent Robots and Automation*, vol. 23, no. 6, pp. 1141-1150, 2007.

[17] C. Ribeiro, "Reinforcement learning agents," *Artificial Intelligence Review*, vol. 17, no. 3, pp. 223-250, 2002.

[18] W. D. Smart and L. Kaelbling, "Practical reinforcement learning in continuous spaces," *in Proc. of the 17th Int. Conf. on Machine Learning*, pp. 903-910, 2000.

[19] C. J. C. H. Watkins and P. Dayan, "*Q*-learning," *Machine Learning*, vol. 8, pp. 279-292, 1992.

[20] P. Y. Glorennec, "Reinforcement learning: an overview," *European Symposium on Intelligent Techniques*, Aachen, Germany, 2000.

[21] Y. Wang, M. Huber, V. N. Papudesi, and D. J. Cook, "User-guided reinforcement learning of robot assistive tasks for an intelligent environment," *in Proc. of the IEEE/RJS Int. Conf. on Intelligent Robots and Systems*, vol. 1, pp. 424-429, 2003.

[22] U. Kartoun, H. Stern, and Y. Edan, "Bag classification using support vector machines," *Applied Soft Computing Technologies: The Challenge of Complexity Series: Advances in Soft Computing*, Springer Berlin / Heidelberg, pp. 665-674, 2006.

[23] U. Kartoun, "Human-Robot Collaborative Learning Methods," Ph.D. dissertation, Dept. of Industrial Engineering and Management, Ben-Gurion University of the Negev, 2007.